Collocations et traitement automatique des langues

Patrick Watrin

Collocations et traitement automatique des langues.

Patrick Watrin [1]
IGM, Université de Marne-la-Vallée & CNRS

Abstract

Nous présentons ici une expérience visant à extraire automatiquement, au départ d'un corpus spécialisé ou non, les unités polylexicales pour ensuite les organiser selon un formalisme exploitable au sein d'une procédure d'analyse syntaxique automatique. Par *unité polylexicale* nous entendons toute séquence de mots constituant une unité d'un point de vue sémantique. Nous rapprochons cette notion de celle de *collocation* qui désigne une suite de mots statistiquement significative et syntaxiquement définie dont le sens n'est pas toujours compositionnel.

Keywords : Collocations, Traitement automatique des langues.

1. Introduction

L'hypothèse que nous défendons ici est la suivante : la collocation peut s'apprécier en tant qu'*unité distributionnelle* et, par conséquent, participer, au niveau lexical, de l'idée harrissienne d'une analyse syntagmatique par regroupements successifs (*cf.* (Harris 1964)). En d'autres termes, nous voulons partir des unités lexicales simples pour obtenir un ensemble de structures complexes, les collocations [1]. Ces structures, résultats d'un attachement au niveau lexical, sont ensuite utilisées, au niveau syntaxique, pour atteindre l'articulation en constituants. Notre objectif est donc de simplifier l'attachement syntagmatique en évacuant une partie de sa complexité au niveau lexical.

Cette idée nous amène à définir un niveau d'abstraction intermédiaire aux plans lexical et syntaxique de l'analyse syntagmatique. En ce sens, elle vient compléter la notion de *chunk* développée par (Abney 1991) en nous permettant de travailler, au niveau syntaxique, depuis un ensemble réduit, et donc moins ambigu, d'éléments terminaux. Par conséquent, notre procédure se pose dans la continuité de l'étiquetage et permet, dans le cas d'une annotation par dictionnaire, d'évacuer une partie conséquente des ambiguïtés lexicale et structurale sur base du contexte local (tout comme le ferait un étiqueteur probabiliste). Nous réduisons en effet une séquence d'entrées potentiellement ambiguës en une unique entrée désambiguïsée. Considérons les exemples suivants :

(1) CFF Recycling a ainsi retrouvé au 30 septembre 2003 une **marge d'exploitation** de 3,8% et une marge nette de 1,7%.
(2) Le groupe prévoit pour l'exercice 2003-2004, sur 15 mois, un **chiffre d'affaires net** consolidé de l'ordre de 30 Millions d'Euros et une rentabilité nette de 7% environ.

[1] IGM, Université de Marne-la-Vallée & CNRS, patrick.watrin@univ-mlv.fr

[1] Précisons que de nombreux travaux, avant le nôtre, abordent cette idée d'une réduction de la complexité lexicale par la prise en compte de structures complexes. L'ouvrage le plus intéressant à ce sujet est, selon nous, (Gross 1996). Toutefois, parce qu'il limite son étude aux noms composés, la couverture qu'il obtient n'est pas comparable à la nôtre. En étendant le paradigme des structures complexes à la collocation, moins figée, nous sommes à même de multiplier les entrées de manière conséquente.

Les deux séquences mises en évidence au sein des exemples 1 et 2 – *marge d'exploitation* et *chiffre d'affaires net* – génèrent une combinatoire lexicale assez importante, 10 analyses possibles pour la première et 120 pour la seconde. En considérant ces deux séquences en tant qu'unités, nous élaguons la combinatoire globale de manière conséquente. Par ailleurs, les structures lexicales complexes de type *NA*, *NprepN* et leurs dérivés nous permettent d'évacuer, dès le niveau lexical, une part de l'ambiguïté d'attachement adjectival et prépositionnel au nom et, ainsi, de résoudre une part non négligeable du travail de rattachement syntagmatique.

Dans cet article, nous décrirons et évaluerons notre procédure d'identification des collocations. Nous proposerons également un formalisme permettant leur utilisation au sein d'une procédure d'analyse syntaxique automatique.

2. Procédure d'extraction

(Manning & Schütze 1999) définit la collocation comme une suite de mots statistiquement significative. En ce sens, elle n'apparait pas syntaxiquement motivée. Bien souvent, d'ailleurs, l'extraction de candidats termes repose sur l'utilisation d'une simple fenêtre de lecture. En ajoutant un ensemble de contraintes syntaxiques, nous limitons nécessairement le champ des possibles mais, parallèlement, nous réduisons considérablement le bruit induit par les méthodes purement quantitatives. De même que (Daille 1995), nous exprimons ces contraintes sous la forme d'un ensemble de filtres linguistiques (*cf.* Figure 1) qui, associés à la sortie d'un étiqueteur grammatical, nous permettent d'établir une liste de candidats cohérents d'un point de vue syntaxique. Nous utilisons ensuite un moteur d'analyse fréquentielle afin d'estimer la pertinence des différents candidats.

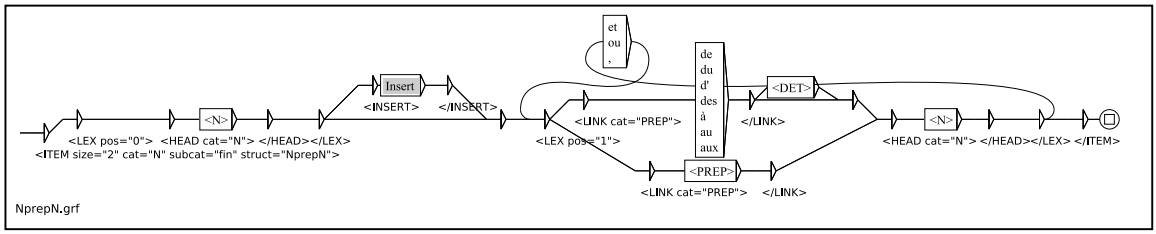


Figure 1. Exemple de filtre linguistique (N prep N)

2.1. Typologie des structures

Dans cette expérience, nous limitons l'extraction à un ensemble de structures nominales de bigrammes (*A-N*, *NA*, *NprepN*, *NN* et *prep-N*) et de trigrammes (*NAA*, *NAN*, *NNA*, *NNprepN*, *NprepNA*, *NprepNN*, *NprepNprepN*). Au-delà de cette taille, les résultats deviennent moins stables et demandent une vérification humaine accrue compromettant l'automatisation du processus.

Le degré de figement des collocations peut fluctuer considérablement. Pour satisfaire à cette propriété nous avons pris en compte, dans nos filtres linguistiques (*cf.* Figure 1), certaines altérations de la structure canonique : la modification et la coordination.

2.1.1. Modification

Les collocations peuvent être augmentées d'un ou plusieurs modifieurs (adjectifs ou adverbes). Par conséquent, selon la complexité, et le degré de figement attendu de la structure, nous permettons ou refusons la présence d'insertions (essentiellement adjectivales), balisées `<INSERT>`.

(3) plate-forme `<INSERT>`centrale`</INSERT>` de traitement et de production
(4) résultat `<INSERT>`brut`</INSERT>` d'exploitation
(5) demandes `<INSERT>`hebdomadaires`</INSERT>` d'allocation chômage

L'une ou l'autre insertion adjectivale forme avec le substantif droit une structure de bigramme de type *NA*. Nous pouvons donc, au départ d'un bigramme (*cf.* Exemples 6 et 7) ou d'un trigramme (*cf.* Exemple 8) sujet à l'insertion, extraire des *ngram* de taille supérieure.

(6) *NprepN* → *NAprepN* : structure <INSERT>financière</INSERT> du groupe

(7) *NprepN* → *NAAprepN* : endettement <INSERT>financier net</INSERT> du groupe

(8) *NprepNprepN* → *NAprepNprepN* : édition <INSERT>spéciale</INSERT> du bulletin interne

2.1.2. Coordination

Les structures *NA* et *NprepN*, dont le degré de figement est moindre, présentent une particularité intéressante, la coordination. Si nous ne la prenions pas en compte, la couverture de notre système d'extraction terminologique ne serait, bien entendu, pas optimale. L'information positionnelle que nous annoncions plus avant nous permet de décompresser la coordination et de créer, au départ d'une extraction unique, autant de candidats que nous avons d'éléments coordonnés (*cf.* (Dominguès 19981999)). La procédure est simple : pour chaque lemme de position *n*, nominal ou adjectival selon le cas, nous créons une entrée en l'associant au lemme de position $n-1$. Ainsi, par exemple, pour le candidat 9, nous créerons les entrées du tableau 1.

(9) <LEX pos="0"> financement </LEX> <LEX pos="1"> de croissance </LEX> <LEX pos="2"> **interne** </LEX> et <LEX pos="2"> **externe** </LEX>

Exemple 9
financement de croissance interne
financement de croissance externe

Table 1. Décompression de la structure coordonnée

2.2. Extraction des motifs

Unitex nous permet d'appliquer nos graphes au texte étiqueté [2] et d'obtenir en sortie un fichier ne comptant que les candidats termes augmentés d'étiquettes XML balisant leurs différentes articulations (*cf.* Figure 1). Une dernière étape nous permet de retenir uniquement le lemme, pour chaque étiquette DELA. L'idée ici est, bien entendu, d'optimiser les calculs fréquentiels ultérieurs. Le lemme nous permet en effet d'unifier les différentes flexions d'une même forme.

(10) <ITEM size="2" cat="N" subcat="fin" struct="NprepN"><LEX pos="0"><HEAD cat="N"> {bénéfice,bénéfice.N}</HEAD> </LEX> <INSERT> {net,net.A}</INSERT> <LEX pos="1"> <LINK cat="PREP">{avant,avant.PREP}</LINK><HEAD cat="N">{survaleurs,survaleur.N}</HEAD> </LEX> </ITEM>

Au sein de notre formalisme XML, la balise <ITEM> délimite les frontières des occurrences. Ce premier élément, racine de chaque unité, possède quatre attributs :

- `size` : désigne la taille du *ngram* extrait,
- `cat` : renseigne la catégorie grammaticale du *ngram*,
- `subcat` : renseigne l'information de sous-catégorisation du *ngram*,
- `struct` : décrit la structure syntaxique interne du *ngram*.

[2] Le texte, étiqueté par le *TreeTagger* (*cf.* (Schmid 1994)) est transformé en une suite d'entrées de type DELA pour permettre à *Unitex* d'accéder au lemme et à la catégorie grammaticale de chaque mot sans devoir appliquer les dictionnaires, sources d'ambiguïté.

Les informations de taille, de catégorie et de structure interviennent dans le comptage fréquentiel que nous décrivons au point suivant. Le dernier attribut, la sous-catégorisation, est, quant à lui, utilisé au niveau de la formalisation des résultats (*cf.* Section 3.) et nous permet de spécifier une étiquette sémantique éventuelle. Cet attribut n'a de sens que si nous travaillons dans un domaine de spécialité restreint (*e.g.* la finance, la médecine, *etc.*).

Chaque élément `<ITEM>` contient plusieurs éléments `<LEX>` délimitant les frontières des différentes articulations de la structure complexe. Il requiert un unique attribut `pos` qui donne sa position au sein de la structure. Cette information qui peut paraître redondante avec l'attribut `size` de l'élément `<ITEM>` est directement utile dans le cas d'élément coordonnés.

Les éléments `<LEX>` pouvant, eux-mêmes, être complexes, la balise `<HEAD>` nous permet de cibler plus précisément les têtes lexicales du *ngram*. Un attribut `cat` est associé à chaque élément `<HEAD>` et indique la catégorie grammaticale de la tête. De plus, dans le cas d'une structure prépositionnelle, l'élément `<LEX>` peut également contenir une balise `<LINK>` dont la fonction est d'identifier et de catégoriser (attribut `cat`) le lien prépositionnel. Notons toutefois que la structure $prep-N$ n'utilise pas la balise `<LINK>` afin d'isoler la préposition. Nous considérons, ici, la préposition comme le premier élément tête de la structure.

2.3. Estimation fréquentielle

Les différentes têtes des structures extraites (balisés `<HEAD>`, dans nos graphes) forment les *ngrams* pour lesquels nous effectuons une analyse fréquentielle. Cette analyse est double, à la fois interne et externe. Pour chaque *ngram*, nous calculons la force d'association de ses composantes à l'aide du log-likelihood. De nombreuses études consacrées aux mesures statistiques ((Dunning 1993), (Manning & Schütze 1999), (Daille 1995)) présentent, en effet, le *log-likelihood* comme la solution la plus adaptée à l'étude des collocations.

L'étude du *log-likelihood* repose sur la création d'une table de contingence exprimant les différentes fréquences associées à chaque paire (L_i, L_j) :

	L_j	$L_{j'}$ $(j' \neq j)$
L_i	a	b
$L_{i'}$ $(i' \neq i)$	c	d

où,

- a est la fréquence des *ngrams* comprenant à la fois L_i et L_j ;
- b est la fréquence des *ngrams* comprenant à la fois L_i et $L_{j'}$;
- c est la fréquence des *ngrams* comprenant à la fois $L_{i'}$ et L_j ;
- d est la fréquence des *ngrams* comprenant à la fois $L_{i'}$ et $L_{j'}$;
- $a+b+c+d$ est le nombre total (N) d'occurrences d'un motif donné.

Nous créons cette table de contingence pour chaque *ngram* obtenu après extraction des candidats termes. Les valeurs a, b, c et d sont ensuite injectées dans la formule du *log-likelihood* telle qu'elle est proposée par (Daille 1995) :

$$\begin{aligned} LL_{(L_i,L_j)} = \; & a\log a + b\log b + c\log c + d\log d \\ & -(a+b)\log(a+b) - (a+c)\log(a+c) \\ & -(b+d)\log(b+d) - (c+d)\log(c+d) \\ & +(a+b+c+d)\log(a+b+c+d) \end{aligned}$$

Le *log-likelihood* est une loi binomiale. Dès lors, son utilisation se limite au test d'hypothèses sur des *ngram* où $n = 2$, c'est-à-dire des bigrammes. Nous ne pouvons donc prétendre analyser de la sorte la force d'association entre les différents membres d'un *ngram* pour $n > 2$.

Nos structures de trigrammes offrent toutefois une perspective intéressante. En découpant un trigramme donné de part et d'autre de l'élément central, nous obtenons une suite de deux bigrammes (*e.g.* $NprepNA \rightarrow NprepN + NA$). Nous pouvons donc envisager l'estimation fréquentielle d'un trigramme et, par transitivité, de tout *ngram*, au départ des bigrammes qu'il contient. Néanmoins, nous ne pouvons simplement additionner les résultats des bigrammes. Un bigramme dont le score d'association est élevé pourrait, en effet, compromettre l'estimation du *ngram* total en lui conférant un score élevé quels que soient les éléments qui suivent. Pour pallier ce problème, (Seretan *et al.* 2003) propose la moyenne statistique suivante :

$$ILL_{(H_0,H_1,\ldots,H_n)} = \frac{(n-1)\prod_{i=0}^{i<n-1} LL(H_i,H_{i+1})}{\sum_{i=0}^{i<n-1} LL(H_i,H_{i+1})}$$

où $LL(H_i,H_{i+1})$ correspond à la mesure de *log-likelihood* pour deux éléments successifs du *ngram*.

La validation statistique peut augmenter la mesure de ***co-occurrence*** que nous venons d'exposer, d'une mesure de ***récurrence***. Cette dernière n'a d'intérêt que si nous travaillons dans un domaine de spécialité. Elle nous permet de contraindre davantage les résultats en estimant l'adéquation d'un candidat au domaine traîté.

La méthode est à nouveau très simple à mettre en œuvre. Au départ d'un corpus de référence et d'un corpus de spécialité, nous créons deux listes de fréquences. Les fréquences de chaque *ngram*, au sein des deux corpus, nous permettent de remplir la table de contingence suivante :

	CORPUS S	CORPUS R
F_n	a	b
$!F_n$	c	d

où,

- *a* est la fréquence du *ngram* de structure S au sein du corpus de spécialité ;
- *b* est la fréquence du *ngram* de structure S au sein du corpus de référence ;
- *c* est la fréquence totale de *ngram* de structure S dans le corpus de spécialité moins la fréquence du *ngram* étudié ;
- *c* est la fréquence totale de *ngram* de structure S dans le corpus de référence moins la fréquence du *ngram* étudié.

Le *log-likelihood*, tel qu'il est précisé par (Rayson & Garside 2000), repose d'une part sur les valeurs *observées* du *ngram*, notées *a* et *b* au sein de la table de contingence et, d'autre part, sur les valeurs *attendues* du *ngram*. Ces dernières se calculent à l'aide des deux formules suivantes et permettent ensuite de résoudre l'équation générale :

$$E_S = \frac{((c+a)*(a+b))}{(c+a+d+b)}$$
$$E_R = \frac{((d+b)*(a+b))}{(c+a+d+b)}$$

$$LL = 2*((a*\log\frac{a}{E_S})+(b*\log\frac{b}{E_R}))$$

Il ne s'agit pas ici de valider la vraisemblance du *ngram* mais plutôt son appartenance à une langue de spécialité donnée. Dès lors, un *ngram* n'est pas considéré comme une suite de *n* unités mais comme une unique entité. Par conséquent, même si la valeur de *n* est supérieure à deux, le *log-likelihood* proposé ci-dessus reste opérationnel.

2.4. Evaluation

Pour évaluer la précision de notre extracteur, nous l'avons confronté à un corpus de 10 000 dépêches du site `firstinvest.com` [3] (1 934 424 mots). Nous avons ensuite vérifié les 8 000 premières entrées de bigrammes et de trigrammes ce qui, par rapport à la totalité des structures identifiées, équivaut à 22,5% des bigrammes différents (*i.e.* les *types* [4]) et à 34% des trigrammes différents. Notons, que ces 8 000 premières entrées couvrent déjà 65,5% des co-occurrences totales de bigrammes (*i.e.* les *tokens*) et 46,56% des co-occurrences de trigrammes.

En termes de précision, les 8 000 entrées considérées offrent **91,63%** de pertinence pour les bigrammes et **84,93%** pour les trigrammes.

Les erreurs que nous avons pu observer sont majoritairement dues à des phénomènes extérieurs à l'estimation fréquentielle qui représentent 52,24% du nombre total d'erreurs pour les bigrammes et 53,98% pour les trigrammes. Ces erreurs se répartissent en deux catégories selon qu'elles ressortissent à la qualité graphique du corpus (*cf.* Exemple 11) ou à l'étiquetage préalable à l'extraction (*cf.* Exemple 12). Les erreurs dues au *TreeTagger* sont très nombreuses et représentent 80% des erreurs extérieures à l'estimation fréquentielle pour les bigrammes et 91,8% pour les trigrammes.

(11) *NN*. Celle-ci est le résultat du ralentissement macro-économique observé en Europe et qui n'est que partiellement compensé par le **maintient** de l'activité aux Etats-Unis (53% du chiffre d'affaires) et dans la **zone zone** Pacifique-Asie.

(12) *NN*. Dans le même temps, les Activités diversifiées voient leurs ventes **{reculer, reculer. N}** de 13,6%, à 1,9 million d'euros.

Les erreurs imputables à la procédure d'extraction proprement dite relèvent majoritairement d'un mauvais attachement et représentent 43,75% des erreurs *internes* pour les bigrammes et 51,92% pour les trigrammes (*cf.* Exemple 13). Les erreurs restantes mettent en cause soit des structures de bigrammes ou trigrammes incomplètes (*cf.* Exemple 14), soit une méprise quant à une potentielle coordination (*cf.* Exemple 15), soit, finalement, certains modifieurs (*cf.* 16).

(13) *NprepN*. MSN Premium, qui est encore testé, sera lancé officiellement avant la fin de l'**année au prix** de 9,95 dollars par mois, soit environ 9,2 euros.

(14) *NprepN*. Saint-Gobain (-1,2%) : le groupe annonce la création d'une société commune détenue à parité avec l'américain Owens Corning pour construire une **usine de fils** de verre pour le renforcement.

(15) *NprepN*. Le groupe table sur une croissance de l'ordre de 10% de **son bénéfice** net par action et **de son bénéfice** d'exploitation pour l'ensemble de l'exercice 2004 et sur une hausse de 4% environ de ses ventes, hors effets de change.

(16) *NprepN*. La **manie** `<INSERT>` **aiguë associée** `</INSERT>` **au trouble** bipolaire est une maladie mentale qui touche 3 à 4% de la population adulte.

3. Un dictionnaire de collocations

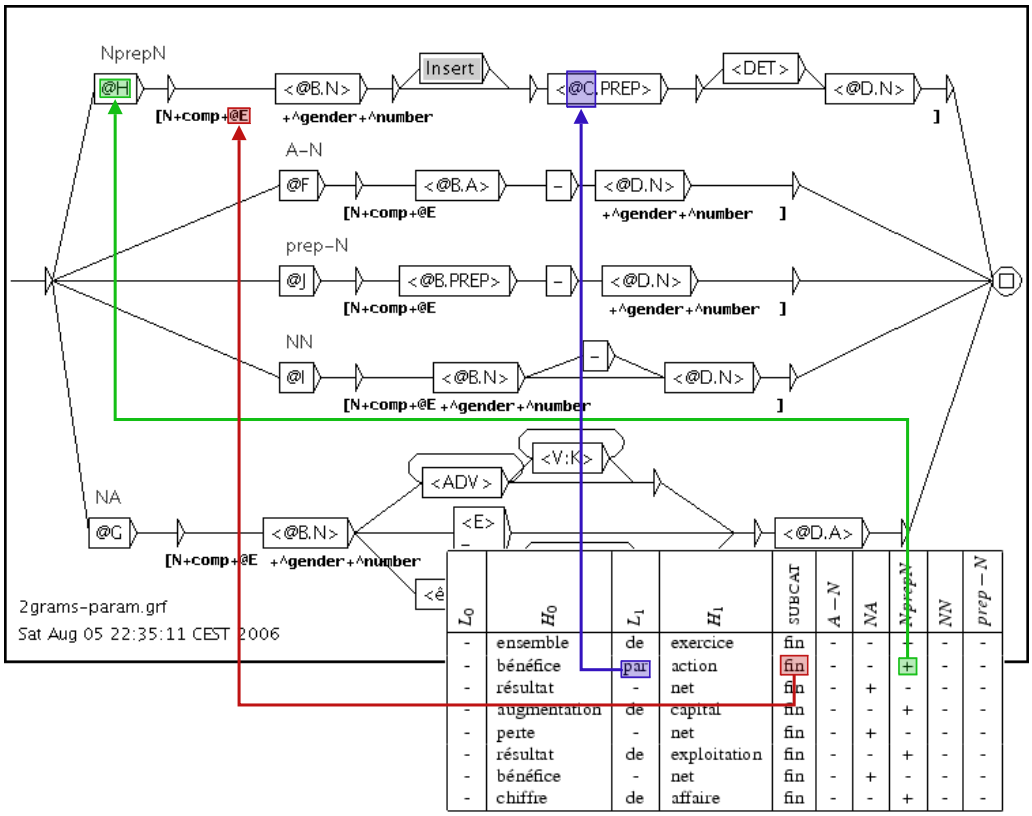
L'extraction terminologique que nous venons de détailler n'est pas une fin en soi. Nous visons avant tout l'amélioration de l'analyse lexicale. Dès lors, se pose la question de la réutilisation des ressources. Pour garantir cette réutilisation, nous devons impérativement valider deux conditions :

[3] Pour cette évaluation, nous travaillions dans un domaine de spécialité homogène (*i.e.* la finance). Nous avons donc pu doublement valider les candidats (mesures de co-occurrence et de récurrence). Afin de permettre la validation de la récurrence, nous avons utilisé un corpus de référence tiré de l'année 2 000 du quotidien belge *Le Soir* (6 308 625 mots).

[4] Pour un corpus donné, les *tokens* désignent l'ensemble des formes identifiées (ici, les bigrammes et trigrammes) et, les *types*, l'ensemble des formes différentes. Le *ratio* des deux nous donne quant à lui une bonne idée de la diversité lexicale d'un corpus.

- **généralisation**, afin de prendre en compte d'éventuelles variations (modifieurs) ;
- **automatisation**, pour permettre l'inclusion du processus au sein d'une chaîne de traitement complexe et non supervisée.

En sortie du processus d'extraction, nous obtenons une suite de *ngrams* formalisés au sein d'une matrice lexico-syntaxique (*cf.* Figure 2). D'un point de vue formel, cette matrice s'apparente à une table du lexique-grammaire : chaque ligne correspond à une entrée lexicale complexe et chaque colonne représente une propriété lexico-syntaxique (*e.g.* la structure syntaxique du collocat et son trait de sous-catégorisation). Aux intersections ligne-colonne, un signe $+$, respectivement $-$, signifie que l'entrée actualise, respectivement n'actualise pas, la structure correspondante.



L_0	H_0	L_1	H_1	SUBCAT	$A-N$	NA	$+NprepN$	NN	$prep-N$
-	ensemble	de	exercice	fin	-	-	+	-	-
-	bénéfice	par	action	fin	-	-	+	-	-
-	résultat	-	net	fin	-	+	-	-	-
-	augmentation	de	capital	fin	-	-	+	-	-
-	perte	-	net	fin	-	+	-	-	-
-	résultat	de	exploitation	fin	-	-	+	-	-
-	bénéfice	-	net	fin	-	+	-	-	-
-	chiffre	de	affaire	fin	-	-	+	-	-

Figure 2. Graphe paramétré

Nous exploitons le contenu des tables au moyen de *graphes paramétrés* (*cf.* (Paumier 2003) et (Roche 1993)). Rappelons que ces graphes formalisent les constructions linguistiques pertinentes pour une table donnée. Chaque construction, c'est-à-dire les structures des différents collocats, correspond à un parcours du graphe (*cf.* Figure 2). Tout comme pour les filtres que nous utilisons lors de l'extraction, nous conservons les points d'insertion.

Cette procédure a été implémentée au sein d'un chunker, décrit dans (Watrin 2006) et (Blanc *et al.* 2007). Les graphes sont appliqués au moyen d'Outilex (Blanc *et al.* 2006). Afin de ne pas perdre toute l'information associée aux entrées qui composent la collocation, nous faisons remonter, au niveau de la structure complexe, le genre et le nombre de la tête lexicale en plus de la catégorie grammaticale. Par ailleurs, chaque structure est augmentée de deux propriétés distributionnelles. La première, $+comp$ (pour complexe), caractérise la structure de l'entrée. La seconde relève, quant à elle, du domaine de spécialité : trait $+fin$ pour la finance, $+med$ pour la médecine, *etc.* Ces différents éléments sont utilisés afin de contraindre l'attachement en syntagmes.

4. Conclusion

Dans cet article, nous avons présenté une méthode fonctionnelle d'extraction d'unités lexicales complexes (ou collocations) de même qu'un format permettant leur utilisation au sein d'un processus d'analyse syntaxique.

La méthode d'extraction offre des résultats tout à fait acceptable rendant possible l'automatisation de la procédure : 91,63% de précision pour les bigrammes et 84,93% pour les trigrammes. Le format d'expression des collocations mis au point est, quant à lui, exploitable et a déjà été intégré au sein d'une procédure de chunking (*cf.* (Watrin 2006) et (Blanc *et al.* 2007), pour plus de détails à ce sujet).

Bien que les mesures que nous utilisons semblent perdre de leur efficacité pour des unités de plus de trois éléments *tête*, nous étudions actuellement la possibilité d'étendre notre méthode à des unités plus grandes afin de pallier à l'erreur d'incomplétude que nous soulignons dans notre évaluation.

References

ABNEY S. P. (1991), "Parsing by Chunks", in Berwick R. C., Abney S. P. & Tenny C. (Eds), *Principle-Based Parsing: Computation and Psycholinguistics*, Kluwer, Dordrecht : 257–278.

BLANC O., CONSTANT M. et LAPORTE E. (2006), "Outilex, plate-forme logicielle de traitement de textes écrits", in Mertens P., Fairon C., Dister A. & Watrin P. (Eds), *Verbum ex machina. Actes de la* 13^{e} *conférence sur le traitement automatique des langues naturelle (TALN'06)*, vol. 1,(Cahiers du Cental) : Presses universitaires de Louvain : 83–92.

BLANC O., CONSTANT M. et WATRIN P. (2007), "Segmentation en super-chunks", in Hathout N. & Muller P. (Eds), *Actes de la* 14^{e} *conférence sur le traitement automatique des langues naturelle (TALN'07)* : IRIT Press : 33–42.

DAILLE B. (1995), *Combined approach for terminology extraction: lexical statistics and linguistic filtering*, Technical report, Lancaster University.

DOMINGUÈS C. (1998/1999), "Traitement de la coordination à l'intérieur des groupes nominaux", in *LinguisticæInvestigationes*, vol. 22.

DUNNING T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", in *Computational Linguistics*, n^{o} 1, vol. 19.

GROSS G. (1996), *Les expressions figées en français. Noms composés et autres locutions*, Ophrys, Paris.

HARRIS Z. S. (1964), "From Morpheme to Utterance", in *Language*, vol. 22.

MANNING C. D. et SCHÜTZE H. (1999), *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge.

PAUMIER S. (2003), *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, PhD thesis, Université de Marne-la-Vallée.

RAYSON P. et GARSIDE R. (2000), "Comparing corpora using frequency profiling", in *Proceedings of the workshop on Comparing Corpora* : 1–6.

ROCHE E. (1993), "Une représentation par automate fini des textes et des propriétés transformationnelles des verbes", in *LinguisticæInvestigationes*, n^{o} 1, vol. 17.

SCHMID H. (1994), "Probabilistic Part-of-Speech Tagging Using Decision Trees", in *International Conference on New Methods in Language Processing*,Manchester, UK.

SERETAN V., NERIMA L. et WEHRLI E. (2003), "Extraction of Multi-Word Collocations Using Syntactic Bigram Composition", in *Proceedings of th* 4th *International Conference on Recent Advances in NLP (RANLP-2003)* : 424–431.

WATRIN P. (2006), *Une approche hybride de l'extraction d'information : sous-langages et lexique-grammaire*, PhD thesis, Université catholique de Louvain.